G3, 2022, 12(5), jkac059
https://doi.org/10.1093/g3journal/jkac059
Advance Access Publication Date: 14 March 2022
Software and Data Resources

GeneCup: mining PubMed and GWAS catalog for gene–keyword relationships

Mustafa H. Gunturkun [1], Efraim Flashner,[2] Tengfei Wang [1], Megan K. Mulligan [2], Robert W. Williams [2], Pjotr Prins [2] and Hao Chen [1,*]

[1]Department of Pharmacology, Addiction Science and Toxicology, University of Tennessee Health Science, Memphis, TN 38103, USA
[2]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science, Memphis, TN 38103, USA

*Corresponding author: Department of Pharmacology, Addiction Science and Toxicology, University of Tennessee Health Science Center, Room 209 Translational Science Building, 71 South Manassas Street, Memphis, TN 38103, USA. Email: hchen@uthsc.edu

Abstract

Interpreting and integrating results from omics studies typically requires a comprehensive and time consuming survey of extant literature. GeneCup is a literature mining web service that retrieves sentences containing user-provided gene symbols and keywords from PubMed abstracts. The keywords are organized into an ontology and can be extended to include results from human genome-wide association studies. We provide a drug addiction keyword ontology that contains over 300 keywords as an example. The literature search is conducted by querying the PubMed server using a programming interface, which is followed by retrieving abstracts from a local copy of the PubMed archive. The main results presented to the user are sentences where gene symbol and keywords co-occur. These sentences are presented through an interactive graphical interface or as tables. All results are linked to the original abstract in PubMed. In addition, a convolutional neural network is employed to distinguish sentences describing systemic stress from those describing cellular stress. The automated and comprehensive search strategy provided by GeneCup facilitates the integration of new discoveries from omic studies with existing literature. GeneCup is free and open source software. The source code of GeneCup and the link to a running instance is available at https://github.com/hakangunturkun/GeneCup.

Keywords: literature mining; PubMed; web service; addiction; custom ontology

Introduction

We describe a web service and application—*Mining gene relationships using custom ontology from PubMed* (GeneCup) (http://genecup.org)—that automatically extracts information from PubMed on the relationship of any gene with a list of user-provided keywords that are hierarchically organized into an ontology. In addition, genetic associations related to the keywords are retrieved from the NHGRI-EBI GWAS catalog. As an example, we created an ontology for drug addiction-related concepts containing 7 categories and over 300 keywords. We describe the details of GeneCup by using this ontology.

Omic studies are becoming the main driving force for discovering molecular mechanisms of human diseases. Over 5,000 genome-wide association studies (GWAS) have mapped over 71,000 associations between genetic variants and diseases/traits (Buniello *et al.* 2019). For example, 1 recent survey identified 1,223 genome-wide significant SNPs associated with psychiatric phenotypes (Horwitz *et al.* 2019). Specialized databases, such as the GWAS catalog (Buniello *et al.* 2019), are available for searching the association between genetic variants and phenotypes. Transcriptome (Farris *et al.* 2015b; Lo Iacono *et al.* 2016; Zhang *et al.* 2016; Cates *et al.* 2019; Kapoor *et al.* 2019; Huggett and Stallings 2020) or epigenome (Ponomarev *et al.* 2012; Farris *et al.* 2015a; De Sa Nogueira *et al.* 2019) profiling using bulk tissue or single cells (Avey *et al.* 2018; Karagiannis *et al.* 2020) have also discovered the involvement of many genes in response to drugs of abuse, stress, or other psychiatric related conditions. Studies using model organisms, such as worms, flies, mice, and rats, have also identified many associations between genetic variants and drug abuse-related phenotypes (Engleman *et al.* 2016; Adkins *et al.* 2017; Highfill *et al.* 2019; Zhou *et al.* 2019).

In these omics studies, understanding the function of genes is a challenging task that requires thorough integration of existing knowledge. Statistics-driven gene ontology, or pathway analysis, are often employed for this purpose. However, an extensive review of the primary literature is ultimately needed to provide a comprehensive and nuanced narrative of these mechanisms. For many scientists, this starts with searches of PubMed based on their domain knowledge. These *ad hoc* searches often miss important information not only because of the inherent complexity of the biology, but also because of the amount of time required for designing a search strategy, conducting the searches, reading the texts, extracting relevant facts, and organizing them into categories. The task of literature searches is especially daunting when many genes are identified in a single study.

Many applications and software were designed for text mining in PubMed (Wei *et al.* 2016). For example, PubTator

Received: January 18, 2022. **Accepted:** March 04, 2022

Central (Wei *et al.* 2019) is a web-based system that makes automatic annotations (i.e. color code genes, diseases, chemicals, etc.) in PubMed for biomedical concepts. Similarly, Thalia (Soto *et al.* 2019) is a semantic search tool that automatically tags concepts occurring in articles indexed in PubMed. We developed Chilibot (Chen and Sharp 2004), which extracts gene–gene or gene–keyword relationships from PubMed. NCBI also maintains GeneRIF (Jimeno-Yepes *et al.* 2013), a database of sentences that describes gene functions. GeneCup fills the void as a web service that extracts relationships between genes and a list of keywords from the entire collection of PubMed abstracts, as well as gene to phenotype associations from the GWAS catalog. GeneCup was designed to meet our own needs of understanding the results from GWAS on addiction-related traits but our web service also allows users to create their own keyword ontologies for a field of interest.

GeneCup relies mostly on keyword matching to select relevant sentences. However, as in the example of addiction ontology the same keyword can have multiple meanings. In particular, stress promotes initial drug use, escalates continued drug use, precipitates relapse and is a major factor contributing to drug addiction (Koob and Schulkin 2019). Stress in this context refers to the body's response to internal and external challenges and is mediated by activating the hypothalamic–pituitary–adrenal axis. In addition, stress can also refer to the responses of cells to perturbations of their environment, such as extreme temperature, mechanical damage, or accumulation of metabolites, etc. These responses often involve the activation of specific molecular pathways. Both systemic and cellular stress have a large collection of literature and it is useful to automatically group sentences containing the word stress according to their exact meaning. There are many machine learning methods that have been applied to this type of natural language processing (NLP) tasks. Young *et al.* (2018) compared some of the deep learning-related algorithms employed in NLP tasks. Among them, convolutional neural network (CNN) has been shown to be efficient in many sentence-level classification projects (dos Santos *et al.* 2015; Francis-Landau *et al.* 2016; Lopez and Kalita 2017; Gehring *et al.* 2017; Wang and Gang 2018). CNN was initially designed for 2D image processing (Lecun *et al.* 1998). It uses a linear operation called convolution besides the regular neural network components, and explores the important patterns in a data by identifying both local and global features of the data. The ability to detect nonlinear relationships among the features effectively is one of the key advantages of deep learning architectures. We therefore developed a CNN to separate sentences describing cellular stress from those that describe system stress.

GeneCup is available as a free web-service. In addition, its source code is available for those interested in setting up a service of their own or modifying the code to better suit their needs.

Methods

System overview

GeneCup is a free and open source web application (Fig. 1). The source code and URL of a running instance is available at https://github.com/hakangunturkun/GeneCup. The main user interface contains a search box that accepts up to 200 gene symbols from the user. Because GeneCup does not search for relationships between gene symbols provided by the user, lists with more genes can be broken into multiple searches without affecting the results. Each gene symbol is paired with each one of the custom ontological categories to query PubMed. The title and abstract of these records are then obtained from a mirrored copy of PubMed on the local server. Sentences containing at least 1 gene symbol and 1 keyword are retained. A local copy of NHGRI-EBI GWAS catalog is also searched for associations between the queried genes and phenotypes related to the ontology. The results are available as an interactive graph or a table that provides links to key sentences from the abstracts, which in turn, are linked to PubMed. In the example of addiction keywords, sentences that contain the keyword "stress" are further classified into 2 types (i.e. systemic vs cellular) before presented to the user, by using a 1D CNN.

We have setup a demonstration server at https://genecup.org. The server is a 28 core Penguin Computing Relion 2600GT system with NVIDIA Tesla K80 GPU. We also tested the software on computers with an Intel i7 CPU with 8 cores and did not notice significant differences in performance. Disk usage is about 122 GB. Most of it is occupied by the local copy of PubMed. Peak memory usage, as estimated by the Resource package in Python, was approximately 420–440 MB during searching, 360 MB for generating the table view, and 440 MB for producing the graphic view. Memory usage during the classification of stress-related sentences using the deep learning module was approximately 4 GB.

Sources of data: PubMed and GWAS catalog

We created a copy of the entire PubMed abstract on our server following instructions provided by the NCBI (Kans 2020). This allows us to rapidly retrieve the abstracts and bypass the limits imposed by NCBI on automated retrievals to prevent system overload. This local copy is updated automatically every week on our server. Downloading the PubMed archive took approximately 106 min over an Internet2 connection. The time required for PubMed synchronization, which is made once a month by a utility of NCBI, is minimal (usually less than 5 min).

We also store a local copy of the GWAS catalog database (Buniello *et al.* 2019) (i.e. all associations v1.0.2 from https://www.ebi.ac.uk/gwas/docs/file-downloads). More specifically, we extract the following fields from the GWAS catalog: PUBMEDID, DISEASE/TRAIT, MAPPED_TRAIT, REPORTED GENES, MAPPED_GENE, SNPS, P-VALUE. The trait fields and gene fields are included in GeneCup searches. All extracted fields are included in the search results. This file is updated manually upon every new release of the catalog.

User-defined ontologies

The custom ontology has 3 levels. The top level is the name of the categories, which can be used to decide whether its sub-categories are included in a new search. The second level is concepts that are displayed in the results (i.e. interactive graphs and tables). The third level contains the actual keywords used in PubMed queries and finding matching sentences. For example, the top level "cells" can contain second level concepts such as "neurons" and "glial cells," with "glial cells" further containing keywords such as "astrocytes," "microglia," etc. The matching keywords at the third level are highlighted using bold font when the sentences are displayed. A special top level keyword "GWAS" is reserved for searching the GWAS catalog. Any keyword under this branch is used to search the GWAS catalog database. This is a flexible structure that allows the user to freely organize a large collection of keywords to fit their needs. A free user account is needed for creating and editing custom ontologies.

As an example, we created a mini-ontology for addiction-related concepts (Supplementary Table 1). The top level has the following 7 categories: addiction stage, drugs, brain region, CNS cell type, stress, psychiatric diseases, and molecular function.

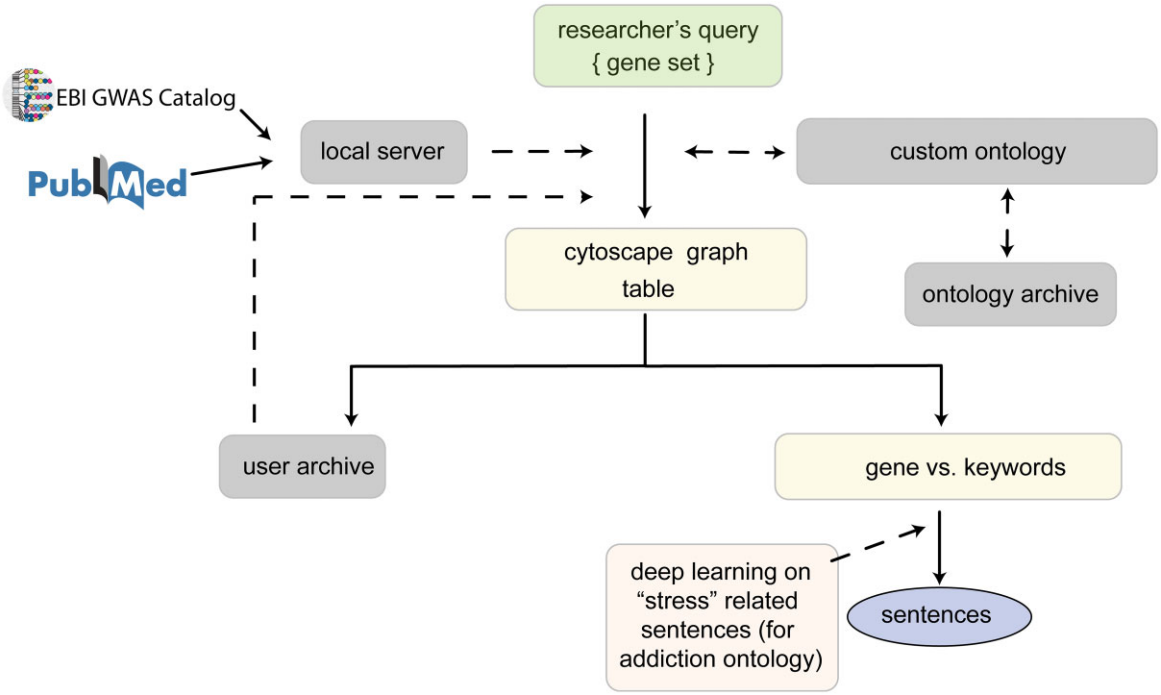

Fig. 1. Overview of the workflow of GeneCup. GeneCup allows users to query the relationship of any gene with a list of keywords hierarchically organized into a custom ontology. This information is automatically extracted from PubMed and NHGRI-EBI GWAS catalog. The users have an option to choose keyword categories during the search. Searches are conducted using EUtils against the PubMed database but abstracts are retrieved from a locally mirrored copy of PubMed. The results are displayed as a cytoscape graph (Fig. 3) and a table. The graph and the table have many interactive elements, including displaying sentences that include the gene symbols and the keywords. Custom ontologies and search results are archived on the server if the user chooses to log in. When the default addiction ontology is used, sentences containing the keyword *stress* are classified using a CNN into 1 of 2 classes: systemic stress or cellular stress (Figs. 2 and 4). Dashed lines: Server operations invoked as needed. Solid lines: Server operations for default queries.

The second level is composed of relevant keywords and the third level includes subconcepts of the keywords or commonly used spelling or acronyms for the keywords. Users have the option to omit any category from the search.

Query processing and user interfaces

We wrote the web-service in the Python programming language and used the Flask library as the web application framework. Users of the web service have the option of creating an account to save search results for later reviews. Query terms provided by the user are first paired with all the keywords. Keywords belonging to the same second level ontology terms are combined using the boolean OR operator before joining with the gene symbol using the AND operator. The E-utilities provided by the NCBI Entrez system (Kans 2020) are used to send the query to the PubMed database (using Esearch) and to retrieve PMIDs (using Efetch). Corresponding records for each PMID are obtained from the local copy of PubMed and the xtract tool is used to parse the titles and abstracts. The Python NLTK library (Bird *et al.* 2009) is then used to tokenize the abstracts into sentences. Python regular expressions are used to find sentences that contain at least one instance of a query gene and 1 instance of a keyword. The number of *abstracts* containing such sentences are then counted. The gene is also searched in the GWAS catalog for phenotypic associations. The number of associations is counted. A network graph is constructed using the Cytoscape Javascript library (Shannon *et al.* 2003), where all genes, keywords, and GWAS terms are used as nodes, and a connection is made between nodes describing a gene and a keyword. The number of abstracts is used as the weight of the edge. This interactive graph allows a user to click on the edge to review the corresponding sentences. All sentences are linked to their original PubMed abstract. The user can also click on a gene to see its synonyms. These synonyms are obtained from the NCBI gene database but are not included in the original search. This is because they often do not appear in the literature or have other meanings and thus provide inaccurate results. However, the web interface allows these synonyms to be included in a new search to retrieve additional information that is potentially relevant.

The GeneCup source code is distributed as free and open source software and can therefore easily be installed on other systems. We have tested our code in commonly used Linux distributions (Debian, GNU Guix, Ubuntu, Arch Linux). The whole service with dependencies is described as a byte reproducible GNU Guix software package (Wurmus *et al.* 2018).

Convolutional neural network to classify sentences describing stress

We trained a 1D CNN to classify sentences describing stress to either cellular stress or system stress (Fig. 2). To create a training corpus, we used a word2vec embeddings library based on PubMed and PubMedCentral data (Moen and Ananiadou 2013) by retrieving words that are similar to examples of systemic stress and cellular stress (e.g. restraint, corticosterone, CRH, and oxidative stress respectively). We then manually crafted 2 PubMed queries to retrieve abstracts related to systemic or cellular stress:

1) (CRF OR AVP OR urocortin OR vasopressin OR CRH OR restraint OR stressor OR tail-shock OR (social AND defeat) OR (foot AND shock) OR immobilization OR (predator AND odor) OR intruder OR unescapable OR inescapable OR CORT

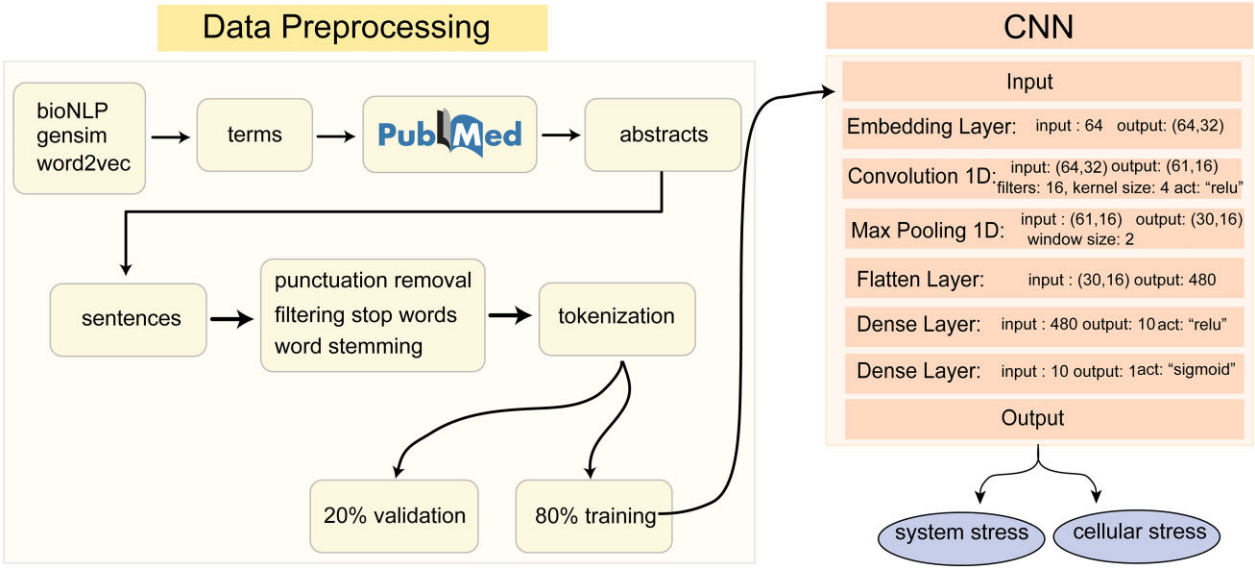

Fig. 2. Pipeline for training the CNN that classifies sentences containing the word "stress." Terms specific to "system stress" or "cellular stress" were obtained by using the cosine similarity tool in Python's Gensim library against the word2vec embeddings derived from PubMed and PMC text. Abstracts including these terms were fetched from PubMed. These words were then "tokenized" and were splitted into training and validation sets. Input layer of the model passed the training data to the embedding layer. After a 1D convolutional layer, downsampling is implemented by a maximum pooling layer. Output is flattened and connected to 2 fully connected layers. We use the rectifier unit function to activate the neurons in the convolution layer and the dense layer. Last dense layer is activated by the sigmoid function. The final weights of the model classify input sentences into either system stress or cellular stress.

OR corticosterone OR cortisol or ACTH OR prolactin OR PRL OR adrenocorticotropin OR adrenocorticotrophin) AND stress NOT (ROS OR oxidative OR redox-regulation OR nitrosative OR nitrative OR hyperglycemia OR carbonyl OR lipoxidative OR Nrf2-driven OR thiol-oxidative)

2) (ROS OR oxidative OR redox-regulation OR nitrosative OR nitrative OR hyperglycemia OR carbonyl OR lipoxidative OR Nrf2-driven OR thiol-oxidative) AND stress

We downloaded all the PubMed abstracts returned from these 2 queries. Manually examining some of the abstracts confirmed the relevance of the results. We then extracted all sentences containing the word *stress* from each set and kept 9,974 sentences from the "systemic stress" class and 9,652 sentences from the "cellular stress" class as our stress training/validation corpus. We maintained another set of 10,000 sentences as the testing corpus, 5,000 sentences for each class.

To clean the data and prepare it for deep learning, we split 19,626 sentences into words, removed punctuation marks, filtered the stop words, and stemmed the words (Brownlee 2017). These words formed a vocabulary of size 23,153 and were tokenized by the Tokenizer library of Keras API. Then the tokenized sentences were split randomly into training and validation sets at 80% and 20%, respectively. We built a 1D CNN in Keras on top of the Tensorflow framework (Abadi *et al.* 2016). The model includes an embedding layer that projects each word to a 32 dimensional space; hence this layer produces a weight matrix with 23,153 × 32 dimensions. Sentences are padded to 64 words, resulting in 64 × 32 sized matrices in the model. After that, a 1D convolutional layer with 16 filters and a kernel size of 4 is implemented and activated by the rectified linear unit (ReLU). This layer produces a 4 × 32 × 16 weight matrix. Downsampling is performed by max pooling with a window size of 2. Then a flattened layer with 480 neurons is connected to 2 fully connected layers, 1 of which has 10 neurons activated with ReLU and the latter one is the final layer activated with a sigmoid function. We validate the model using 3,924 sentences, 1,997 of them belong to the "systemic stress" class, 1,927 sentences belong to the "cellular stress" class. These were selected randomly before training. To minimize the value of the loss function and update the parameters, Adamax optimization algorithm (Kingma and Ba 2014) was used with the parameters of learning rate = 0.002, beta1 = 0.9, beta2 = 0.999. The binary cross entropy loss function is used for this binary classification task. These hyperparameters were optimized using the training corpus.

We used the confusion matrix to evaluate the performance of the classification and summarize the results for the test dataset (Table 1). The rows and the columns of the matrix represent the values for the actual class and predicted class, respectively. The measures of accuracy in the table were calculated by using the values in the table; the number of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). Sensitivity, i.e. the ratio of TP to TP+FN, is the proportion of the systemic stress sentences correctly identified. Specificity, i.e. the ratio of TN to TN+FP, is the ability of the model to identify the cellular stress sentences correctly. Precision is the proportion of the correct systemic stress sentences in the predicted class of systemic stress sentences, and is calculated as the ratio of TP to TP+FP. Accuracy of the model is the proportion of the total number of predictions that are correct, and is calculated as the ratio of TP+TN to all. The performance measures including the area under the ROC curve (sensitivity vs 1-specificity) produced by these values are given in the Results section.

Results

We have written a graphical interface for searching the role genes play in biological systems. One running instance is available at http://genecup.org. A query of 3 terms can be completed in about 20–30 s. The query time increases linearly by the number of

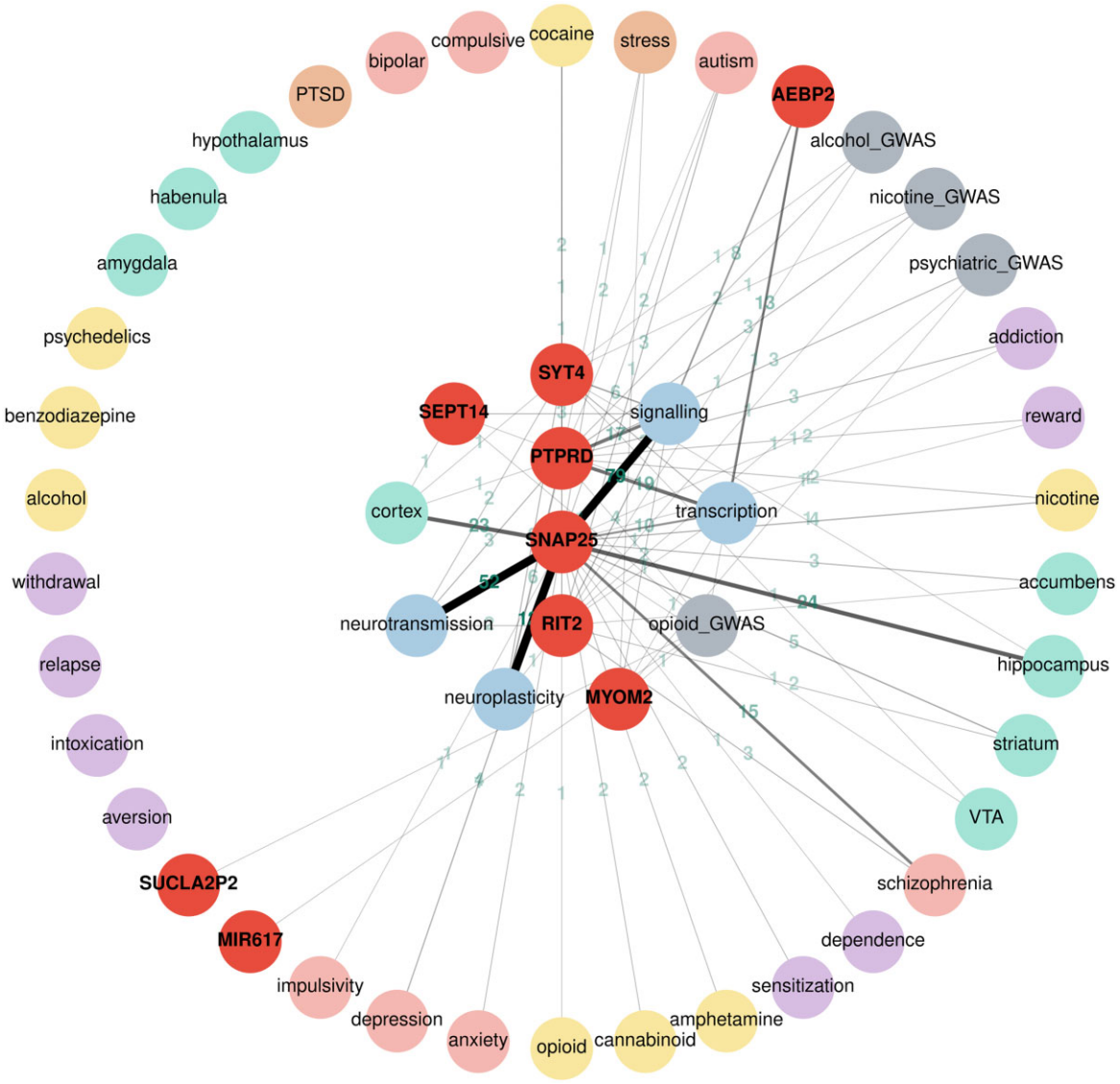

Fig. 3. An interactive Cytoscape graph visualizing gene–keyword relationships. An interactive Cytoscape graph visualizing gene–keyword relationships. Nodes (circles) represent either search terms (in red) or keywords (colored according to the mini ontology; GWAS results are in gray). Clicking the keyword nodes displays the individual terms that are included in the search. Clicking the gene symbols displays their synonyms. The edges represent relationships between nodes. The number of PubMed abstracts where the gene symbol and keyword co-occur in the same sentence are displayed on the edges. The width of the edge is correlated with the number of abstracts. Clicking on the edges shows these sentences, which are linked back to PubMed abstracts. Nodes can be moved about for better visibility of relationships. These genes were taken from a recent genome-wide association study of opioid cessation (Cox *et al.* 2020).

terms. Thus a search of 20 genes can be completed in about 2–3 min. Most of the time is spent on interacting with PubMed to obtain PMIDs. The time required to form a customized ontology for a new subject area is dependent on the familiarity of the subject area. However, entering it into the system requires a minimal amount of time using the custom ontology editor.

In addition, queries can also be initiated by placing the terms in the URL. For example, to start a search for *CHRNA5* and *BDNF* genes against the keyword categories drug, stress, addiction, and GWAS, the following hyperlink can be used: https://genecup.org/progress?type=drug&type=stress&type=GWAS&type=addiction&query=CHRNA5+BDNF

This allows links to GeneCup queries to be embedded into other websites. When the hyperlink above is clicked, the results in graphical format will appear in a separate window.

For sentences containing the word "stress," we designed a 1D CNN with 4 hidden layers (Fig. 2) to differentiate them into 2 classes, namely systemic and cellular stress. The neural network was optimized using the gradient-based optimization algorithm Adamax. During training, model accuracy (Supplementary Fig. 1a) increased rapidly during the first 5 epochs to approximately 0.995, while validation accuracy peaked at 0.991 at epoch five. On the other hand, model loss curve (Supplementary Fig. 1b) on the training dataset continued to decline after the initial drop and approached zero after 15 epochs. However, the loss on the validation data set started to increase after epoch five, indication model overfitting. Therefore, we used the weights that maximized the validation performance before overfitting (i.e. epoch five). By using these weights and parameters, our model has an AUC of 99.2% on the validation dataset.

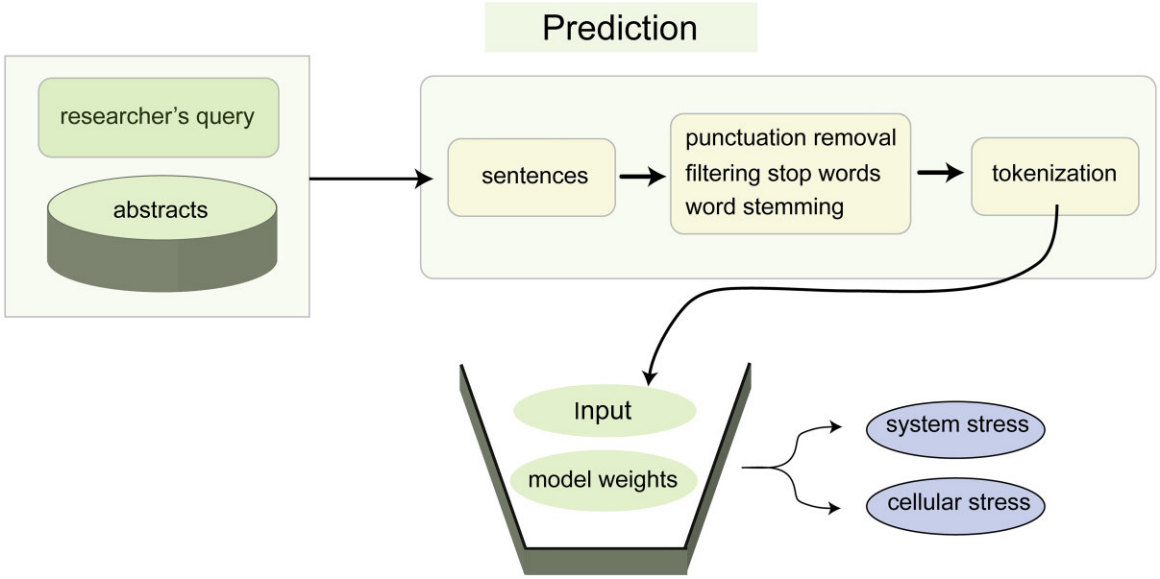

Fig. 4. Steps for classifying sentences using a trained neural network. Steps for classifying sentences using a trained neural network. Abstracts are fetched from the locally mirrored copy of PubMed and are parsed into sentences. Punctuation marks and stop words are removed and the remaining words of the sentences are stemmed. The words are tokenized by using the Tokenizer library of the Keras API. The weight matrices of the trained model are multiplied by the sentence matrix to predict whether the input sentences are related to system stress or cellular stress.

Table 1. Confusion matrix of CNN on test data.

		Predicted class		
		Systemic stress	**Cellular stress**	
Actual class	Systemic stress	4,853 (TP)	147 (FN)	Sensitivity: 97%
	Cellular stress	310 (FP)	4,690 (TN)	Specificity: 94%
		Precision: 94%	Negative predictive value: 97%	Accuracy: 95%

We tested the model on a new dataset consisting of 5,000 system stress sentences and 5,000 cellular stress sentences. The confusion matrix for the prediction of the test dataset is presented as Table 1. The sensitivity of the model (i.e. the proportion of predicted systemic class sentences to all sentences observed in this class) is 97%. The similar measure for cellular class sentences, i.e. specificity is 94%. The prediction accuracy of the model (i.e. the ability to distinguish 2 classes on the test dataset) is 95.4% and the AUC is 98.9% for the test dataset.

We also checked the distribution of the predicted probabilities (Supplementary Fig. 2) of the test dataset. The model predicts a probability of the class membership for each sentence. If the predicted probability of a sentence is more than 0.5, it is labeled as a system stress sentence. Otherwise, the sentence is predicted to be a member of the cellular stress class. Among the system stress sentences in the test dataset, 88% of the sentences had predicted probabilities greater than 0.9. This shows the model's confidence of its prediction on stress sentences. Likewise, 88% of the cellular stress sentences had predicted probabilities less than 0.1. Therefore the model is 90% confident about the classification of 88% of the cellular stress sentences.

The weights of the trained model are saved on the server and are used to make predictions for each retrieved sentence when the user clicks on the edge connecting the stress category and the gene name (Fig. 4). As an example of run time performance, it took approximately 12 s to classify 3,908 sentences on *CRF* and stress.

As a demonstration of the utility of the web interface, we entered the 9 genes that reached suggestive significance in a recent genome-wide association study of opioid cessation (Cox *et al.* 2020). The graph view of the search results are shown in Fig. 3. Genes and keywords are all shown as circles and lines connecting them show the number of abstracts containing the 2 circles they connect. Keywords under the same main category are shown with the same color in the graphic output. Clicking on the lines brings up a new page that displays all sentences containing the keywords that line connects. An alternative tabular view of the same results is also available, where genes, the keywords, and number of abstracts are shown as separate columns.

Our results contained sentences in PubMed that described the roles played by *PTPRD*, *SNAP25*, and *MYOM2* in addiction, which were all discussed in the original publication (Cox *et al.* 2020). In addition, our results found sentences that indicated the potential involvement of *RIT2* and *SYT4* in addiction. For example, *RIT2* is associated with smoking initiation (Liu *et al.* 2019) and autism (Liu *et al.* 2016). Recent publications indicated that *RIT2* is involved in dopamine transporter trafficking (Fagan *et al.* 2020) and plays a sex-specific role in acute cocaine response (Sweeney *et al.* 2020). *SYT4* is expressed in the hippocampus and entorhinal cortex (Crispino *et al.* 1999) and regulates synaptic growth (Harris *et al.* 2016; Ó'Léime *et al.* 2018). Further, *SUCLA2P2* has been implicated in the age of smoking initiation (Argos *et al.*, 2014) and Schizophrenia (Ikeda *et al.* 2019). This example demonstrated the utility of GeneCup in rapidly finding information that links a

gene to addiction and thus integrating new findings with previous research findings.

Discussion

We present here a literature mining web application, GeneCup, that extracts sentences from a locally mirrored copy of PubMed abstracts containing user-provided gene symbols and the keywords of the custom ontology. Associations between the genes and various phenotypes from human GWAS results are also provided. The users can include up to 200 gene symbols in each search. This cutoff number is chosen by considering the completion time of the search. GeneCup does not search the interactions between genes, hence a larger number of terms can be queried in more than 1 round without comprising the results. The results are presented in a graphical or a tabular format, both provide links to review individual sentences that contain the gene and at least 1 keyword. Gene synonyms are also presented and can be included in additional searches. As an example, we provide an ontology containing approximately 300 predefined addiction-related keywords organized into 7 categories. Our ontology editor allows users to modify or create their own ontology to fit their needs. Stress-related sentences are automatically classified into system vs cellular stress if the addiction ontology is used.

Scientists using omics methods face a particularly challenging task when trying to integrate new findings with existing knowledge. The increasing number of genes contained in data sets, the breadth of sciences, and the large amount of existing knowledge captured in PubMed make systematic literature surveys daunting tasks. Typically, scientists manually conduct more detailed searches in areas where they have expertise and the queries are much less thorough in other areas. The search strategies are often crafted ad hoc and likely different from 1 day to another.

Many software or web applications have been created to automatically analyze PubMed abstracts. For example, PubMatrix (Becker *et al.* 2003) reports the frequencies of co-occurrence between 2 lists of terms at the abstract level. Chilibot (Chen and Sharp 2004), which we created, reports sentences that contain gene or keyword relationships extracted from the latest 30–50 abstracts. GeneRIF (Jimeno-Yepes *et al.* 2013) provides sentences on the function of genes but it does not provide an interface for displaying the relationships between multiple query terms. In contrast to these, GeneCup provides an interface that allows comprehensive queries of the function of any gene using a set of user-defined keywords. It searches the entire PubMed database and extracts sentences describing the function of genes. In addition, GeneCup also queries the GWAS catalog, which contains many genetic associations that are not reported in abstracts. Although most of the functions provided by GeneCup can be carried out manually, it will require several orders of magnitude more time and effort. Even then, the manually collected results will be difficult to review. In contrast, results provided by GeneCup are automatically organized by the ontology. All the genes and keywords can be seen in graphs or tables, with informative sentences and abstracts readily available.

GeneCup presents to the user sentences containing genes and keywords of interest to the user. Compared to phrases or abstracts, sentences are the most succinct semantic unit to convey a fact. Ding *et al.* (2002) compared different text processing units for text mining system design and found that the highest precision of information retrieval is achieved when phrases are used as the text unit, whereas using sentences are more effective than both phrases and abstracts. Therefore, similar to Chilibot (Chen and Sharp 2004), we continue to use sentences as the information unit. Unlike the commonly used gene ontology enrichment (Osborne *et al.* 2007) or gene set enrichment (Subramanian *et al.* 2005) analysis, the literature analysis provided by GeneCup does not evaluate any statistical significance. Instead, these key sentences provide easy access to relevant prior research, where the nuanced details can be easily obtained by following the link from the sentence to the abstract and then to the full text article.

Stress plays key roles in addiction but the word stress has multiple meanings. Using a convolutional network, we trained a model that achieved 97% sensitivity and 94% specificity in classifying sentences containing the word stress to either systemic stress or cellular stress. Training such a model requires large amounts of labeled data. Manually labeling these data are very labor intensive. Using an approach that is similar to some recent advances in automated data labeling (Ratner *et al.* 2020), we carefully crafted 2 PubMed queries to obtain over 30,000 sentences that mostly belong to the correct category. This large corpus of text allowed us to achieve peak classification performance with less than 5 epochs of training (Supplementary Fig. 1).

Gene synonyms represent a large challenge to any text mining approach. Not including synonyms will result in the loss of information. However, many synonyms, especially those that are short, have multiple meanings. For example, *CNR* is a synonym for the *CNR1* gene. However, *CNR* is also an acronym for contrast noise ratio, frequently used in imaging analysis literature. For user-supplied gene symbols, we do not include synonyms in the initial search to prevent the noise from "drowning out" the signal. Instead, we provide users an option to either search individual synonyms or to conduct a combined search of all synonyms as a secondary step. We think this middle-of-the-road approach is the more efficient way to achieve a balance between computation and performance. Future work can potentially use deep learning to classify all PubMed abstracts for their relevance to the field of interest and thus exclude many abstracts containing short words that are not relevant.

Many future improvements for GeneCup are possible. For example, GeneCup uses PubMed abstracts as the source of data, rather than PubMed Central, which contains full-text articles. Lin (2009) compared the effectiveness of information retrieval from abstract vs full text search and found that full text search, when indexed using paragraphs as the unit, is more effective than the abstract-only search. Several groups have reported either using full text search for curation (Van Auken *et al.* 2014; Müller *et al.* 2018) or using full text for analysis (Wei and Collier 2011; Verspoor *et al.* 2012; Islamaj Doğan *et al.* 2017). NCBI also provides an API for PubMed Central. However, the majority of the articles in PubMed Central are subject to traditional copyright restriction and it is not feasible to establish a local mirror of the full-text collection. Retrieving text via NCBI API is not feasible on the scale we need (e.g. several thousand articles at a time). Further, we anticipate full text may cause duplications of information and increase the noise in results.

GeneCup currently does not retrieve relationships between genes. There are several existing tools available for this purpose, such as Chilibot (Chen and Sharp 2004), or GeneMania (Warde-Farley *et al.* 2010). Instead, GeneCup focuses on the relationship between genes and a set of keywords organized as an ontology. The addiction ontology was developed based on the expertise of the authors. It certainly contains biases and can be further improved. For example, tight integration with community developed ontology for addiction or psychiatric disease, such as those that are available from the Open Biological and Biomedical

Ontology Foundry (www.obofoundry.org), or automated methods for converting MESH headings can be tested in the future.

Data availability

GeneCup is a free and open source web application. The source code of GeneCup and the link to a running instance is available at https://github.com/hakangunturkun/GeneCup.

Supplemental material is available at G3 online.

Acknowledgments

MHG conducted the research and drafted the manuscript. HC conceived of the project and conducted the initial research. EF, TW, MKM, RWW, and PP contributed to the research. All authors revised and approved the final manuscript.

Funding

Funding is provided by the following NIH/NIDA grants: U01DA047638 (HC and RWW), P30DA044223 (RWW and PP), and also by NIH/NIGMS grant R01GM123489 (RWW and PP).

Conflicts of interest

None declared.

Literature cited

Abadi M, Barham P, Chen J, Chen Z, Davis A. *et al.* TensorFlow: a system for large-scale machine learning, in Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, OSDI'16. USENIX Association, USA, 2016. pp. 265–283.

Adkins AE, Hack LM, Bigdeli TB, Williamson VS, McMichael GO, Mamdani M, Edwards AC, Aliev F, Chan RF, Bhandari P, *et al.*; German Study of the Genetics of Addiction Consortium. Genomewide association study of alcohol dependence identifies risk loci altering ethanol-response behaviors in model organisms. Alcohol Clin Exp Res. 2017;41(5):911–928.

Argos M, Tong L, Pierce BL, Rakibuz-Zaman M, Ahmed A, Islam T, Rahman M, Paul-Brutus R, Rahaman R, Roy S, *et al.* Genome-wide association study of smoking behaviours among Bangladeshi adults. J Med Genet. 2014;51(5):327–333.

Avey D, Sankararaman S, Yim AKY, Barve R, Milbrandt J, Mitra RD. Single-cell RNA-Seq uncovers a robust transcriptional response to morphine by glia. Cell Rep. 2018;24(13):3619–3629.e4.

Becker KG, Hosack DA, Dennis G, Lempicki RA, Bright TJ, Cheadle C, Engel J. PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics. 2003;4:61.

Bird S, Klein E, Loper E. 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Newton, MA, USA: O'Reilly Media Inc.

Brownlee J. 2017. Deep Learning for Natural Language Processing: develop Deep Learning Models for Your Natural Language Problems. Machine Learning Mastery. Retrieved from: https://machinelearningmastery.com/deep-learning-for-nlp/

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–D1012.

Cates HM, Bagot RC, Heller EA, Purushothaman I, Lardner CK, Walker DM, Peña CJ, Neve RL, Shen L, Nestler EJ, *et al.* A novel role for E2F3b in regulating cocaine action in the prefrontal cortex. Neuropsychopharmacology. 2019;44(4):776–784.

Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004;5:147.

Cox JW, Sherva RM, Lunetta KL, Johnson EC, Martin NG, Degenhardt L, Agrawal A, Nelson EC, Kranzler HR, Gelernter J, *et al.* Genome-wide association study of opioid cessation. J Clin Med. 2020;9(1):180. doi:10.3390/jcm9010180.

Crispino M, Stone DJ, Wei M, Anderson CP, Tocco G, Finch CE, Baudry M. Variations of synaptotagmin I, synaptotagmin IV, and synaptophysin mRNA levels in rat hippocampus during the estrous cycle. Exp Neurol. 1999;159(2):574–583.

De Sa Nogueira D, Merienne K, Befort K. Neuroepigenetics and addictive behaviors: where do we stand? Neurosci Biobehav Rev. 2019;106:58–72.

Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: abstracts, sentences, or phrases? Pac Symp Biocomput. 2002;326–337.

dos Santos C, Xiang B, Zhou B. 2015. Classifying relations by ranking with convolutional neural networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: long Papers), Association for Computational Linguistics, Beijing (China). 626–634.

Engleman EA, Katner SN, Neal-Beliveau BS. *Caenorhabditis elegans* as a model to study the molecular and genetic mechanisms of drug addiction. Prog Mol Biol Transl Sci. 2016;137:229–252.

Fagan RR, Kearney PJ, Sweeney CG, Luethi D, Schoot Uiterkamp FE, Schicker K, Alejandro BS, O'Connor LC, Sitte HH, Melikian HE. Dopamine transporter trafficking and Rit2 GTPase: Mechanism of action and in vivo impact. J Biol Chem. 2020;295(16):5229–5244.

Farris SP, Harris RA, Ponomarev I. Epigenetic modulation of brain gene networks for cocaine and alcohol abuse. Front Neurosci. 2015a;9:176.

Farris SP, Arasappan D, Hunicke-Smith S, Harris RA, Mayfield RD. Transcriptome organization for chronic alcohol abuse in human brain. Mol Psychiatry. 2015b;20(11):1438–1447.

Francis-Landau M, Durrett G, Klein D. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: human Language Technologies, Association for Computational Linguistics, San Diego (CA). p. 1256–1261.

Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. 2017. Convolutional sequence to sequence learning. Proceedings of the 34th International Conference on Machine Learning – Vol. 70, ICML'17. JMLR.org. p. 1243–1252.

Harris KP, Zhang YV, Piccioli ZD, Perrimon N, Littleton JT. The postsynaptic t-SNARE Syntaxin 4 controls traffic of Neuroligin 1 and Synaptotagmin 4 to regulate retrograde signaling. Elife. 2016;5:e13881. doi: 10.7554/eLife.13881.

Highfill CA, Baker BM, Stevens SD, Anholt RRH, Mackay TFC. Genetics of cocaine and methamphetamine consumption and preference in *Drosophila melanogaster*. PLoS Genet. 2019;15(5):e1007834.

Horwitz T, Lam K, Chen Y, Xia Y, Liu C. A decade in psychiatric GWAS research. Mol Psychiatry. 2019;24(3):378–389.

Huggett SB, Stallings MC. Cocaine'omics: genome-wide and transcriptome-wide analyses provide biological insight into cocaine use and dependence. Addict Biol. 2020;25:e12719.

Ikeda M, Takahashi A, Kamatani Y, Momozawa Y, Saito T, Kondo K, Shimasaki A, Kawase K, Sakusabe T, Iwayama Y, *et al.* Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. Schizophr Bull. 2019;45(4):824–834.

Islamaj Doğan R, Kim S, Chatr-Aryamontri A, Chang CS, Oughtred R, Rust J, Wilbur WJ, Comeau DC, Dolinski K, Tyers M, *et al.* The BioC-BioGRID corpus: full text articles annotated for curation of protein-protein and genetic interactions. Database. 2017;2017: baw147. doi:10.1093/database/baw147.

Jimeno-Yepes AJ, Sticco JC, Mork JG, Aronson AR. GeneRIF indexing: sentence selection based on machine learning. BMC Bioinformatics. 2013;14:171.

Kans J. 2020. Entrez Direct: e-Utilities on the UNIX Command Line. National Center for Biotechnology Information (US). Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK179288/

Kapoor M, Wang J-C, Farris SP, Liu Y, McClintick J, Gupta I, Meyers JL, Bertelsen S, Chao M, Nurnberger J, *et al.* Analysis of whole genome-transcriptomic organization in brain to identify genes associated with alcoholism. Transl Psychiatry. 2019;9(1):89.

Karagiannis TT, Cleary JP, Gok B, Henderson AJ, Martin NG, Yajima M, Nelson EC, Cheng CS. Single cell transcriptomics reveals opioid usage evokes widespread suppression of antiviral gene program. Nat Commun. 2020;11(1):2611.

Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.

Koob GF, Schulkin J. Addiction and stress: an allostatic view. Neurosci Biobehav Rev. 2019;106:245–262.

Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–2324.

Lin J. Is searching full text more effective than searching abstracts? BMC Bioinformatics. 2009;10:46.

Liu X, Shimada T, Otowa T, Wu Y-Y, Kawamura Y, Tochigi M, Iwata Y, Umekage T, Toyota T, Maekawa M, *et al.* Genome-wide Association Study of Autism Spectrum Disorder in the East Asian Populations. Autism Res. 2016;9(3):340–349.

Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, Datta G, Davila-Velderrain J, McGuire D, Tian C, *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019;51(2):237–244.

Lo Iacono L, Valzania A, Visco-Comandini F, Viscomi MT, Felsani A, Puglisi-Allegra S, Carola V. Regulation of nucleus accumbens transcript levels in mice by early-life social stress and cocaine. Neuropharmacology. 2016;103:183–194.

Lopez MM, Kalita J. Deep Learning applied to NLP. 2017. arXiv [cs.CL]. Retrieved from: arxiv.org/abs/1703.03091

Moen S, Ananiadou TSS. Distributional semantics resources for biomedical text processing. Proceedings of LBM. 2013;39–44.

Müller H-M, Van Auken KM, Li Y, Sternberg PW. Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. BMC Bioinformatics. 2018;19(1): 94.

Ó'Léime CS, Hoban AE, Hueston CM, Stilling R, Moloney G, Cryan JF, Nolan YM. The orphan nuclear receptor TLX regulates hippocampal transcriptome changes induced by IL-1β. Brain Behav Immun. 2018;70:268–279.

Osborne JD, Zhu LJ, Lin SM, Kibbe WA. Interpreting microarray results with gene ontology and MeSH. Methods Mol Biol. 2007; 377:223–242.

Ponomarev I, Wang S, Zhang L, Harris RA, Mayfield RD. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. J Neurosci. 2012;32(5):1884–1897.

Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J. 2020;29(2): 709–730.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504.

Soto AJ, Przybyła P, Ananiadou S. Thalia: semantic search engine for biomedical abstracts. Bioinformatics. 2019;35(10):1799–1801.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–15550.

Sweeney CG, Kearney PJ, Fagan RR, Smith LA, Bolden NC, Zhao-Shea R, Rivera IV, Kolpakova J, Xie J, Gao G, *et al.* Conditional, inducible gene silencing in dopamine neurons reveals a sex-specific role for Rit2 GTPase in acute cocaine response and striatal function. Neuropsychopharmacology. 2020;45(2):384–393.

Van Auken K, Schaeffer ML, McQuilton P, Laulederkind SJF, Li D, Wang S-J, Hayman GT, Tweedie S, Arighi CN, Done J, *et al.* BC4GO: a full-text corpus for the BioCreative IV GO task. Database 2014. 2014. doi:10.1093/database/bau074.

Verspoor K, Cohen KB, Lanfranchi A, Warner C, Johnson HL, Roeder C, Choi JD, Funk C, Malenkiy Y, Eckert M, *et al.* A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics. 2012;13:207.

Wang W, Gang J. 2018. Application of Convolutional Neural Network in Natural Language Processing. 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE). 64–70.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38: W214–W220.

Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1):W587–W593.

Wei C-H, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics. 2016; 32(12):1907–1910.

Wei Q, Collier N. Towards classifying species in systems biology papers using text mining. BMC Res Notes. 2011;4:32.

Wurmus R, Uyar B, Osberg B, Franke V, Gosdschan A, Wreczycka K, Ronen J, Akalin A. PiGx: reproducible genomics analysis pipelines with GNU Guix. Gigascience. 2018;7(12):giy123. doi:10.1093/gigascience/giy123.

Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing. IEEE Comput. Intell. Mag. 2018;13:55–75.

Zhang Y, Kong F, Crofton EJ, Dragosljvich SN, Sinha M, Li D, Fan X, Koshy S, Hommel JD, Spratt HM, *et al.* Transcriptomics of environmental enrichment reveals a role for retinoic acid signaling in addiction. Front Mol Neurosci. 2016;9:119.

Zhou Z, Blandino P, Yuan Q, Shen P-H, Hodgkinson CA, Virkkunen M, Watson SJ, Akil H, Goldman D. Exploratory locomotion, a predictor of addiction vulnerability, is oligogenic in rats selected for this phenotype. Proc Natl Acad Sci USA. 2019;116(26):13107–13115.

Communicating editor: J. M. Cherry